Shrivatsasingh Rathore

Portfolio | +1(480) 703-7417 | shrivrath9@gmail.com | LinkedIn | GitHub | Open to Relocation

Education

MS in Data Science, Analytics and Engineering, Arizona State University — Aug 2024 – May 2026
B.Tech. in Computer Science, Savitribai Phule Pune University — Aug 2020 – Apr 2024

Technical Skills

Languages/Databases: Python, Java, TypeScript, JavaScript, Dart, SQL, PostgreSQL, Redis, Supabase, Bash
Frameworks/Technologies: FastAPI, Flask, Spring Boot, Node.js, Next.js, React, Flutter, Apache Kafka, gRPC, REST APIs, Microservices, Docker, Kubernetes, CI/CD
AI/ML: LLMs, Agentic Systems, RAG, LangChain, Whisper, Ollama, Groq, CNNs, NLP, Computer Vision, Prompting
Cloud/Tools: AWS (EC2, S3, Lambda), Prometheus, Grafana, JUnit 5, Testcontainers, Git, System Design, SQLite, MySQL

Experience

ASU Trustworthy AI Lab

Research Assistant — **May 2026 – Present**

- Investigate **hallucination behavior in radiology vision-language models (LLaVA-Rad)** by developing an image-corruption evaluation pipeline across **10 corruption types** using **MIMIC-CXR** data.
- Improved hallucination-detection **F1 from 0.425 to 0.582** by implementing **multi-sample inference** with 10 samples per image and **CheXbert**-based disease-label extraction, establishing a more robust evaluation baseline.
- Analyze model outputs and failure patterns to characterize **vision-language hallucinations in clinical imaging**, supporting ongoing research into reliable and trustworthy AI for healthcare applications.

Superstars Inc.

Software Developer Intern — **Sep 2025 – Dec 2025**

- Architected backend search APIs with **multi-layer caching** and query optimization, reducing search latency by **35%** across a cross-platform **Flutter/Dart** application serving **thousands of concurrent users**.
- Developed and integrated **Flutter/Dart frontend features** while redesigning API request handling with **debouncing and request deduplication**, eliminating **40% of redundant network calls**.
- Resolved **3 recurring production failures** through structured logging and performance profiling, reducing mean time to resolution by **60%** and eliminating manual on-call escalations.

Dr. D. Y. Patil Unitech Society

Research Assistant — **Aug 2023 – Jun 2024**

- Deployed **AWS infrastructure (EC2, S3, Lambda)** to ingest and preprocess an **80,000-image** plant disease dataset across **15 crop categories**, improving pipeline throughput by **40%** and cutting preprocessing time **3x**.
- Benchmarked **CNNs across 5 architectures** on AWS EC2, achieving **93%+ classification accuracy**; findings published in a **Wiley/O'Reilly book chapter** and **SSRN preprint**.

Projects

VaultStream: Distributed Payment Engine [Github] *Java 21, Spring Boot, Kafka, gRPC, RAFT, PostgreSQL, Redis, Prometheus, Grafana, Docker, Kubernetes*

- Engineered a **fault-tolerant distributed payment engine** using **RAFT consensus over gRPC**, sustaining write availability with majority quorum under node failures; validated through failure injection simulating **2-node simultaneous crashes** across a 3-node Kubernetes cluster.
- Built an **idempotent REST-to-Kafka-to-RAFT-to-PostgreSQL** pipeline with exactly-once delivery, **24-hour Redis deduplication**, Resilience4j circuit breakers with dead-letter fallback, and Prometheus/Grafana observability; achieved **p99 latency under 350ms**.

Reachr: AI Voice Recruiting Platform [Github] *Next.js, FastAPI, Python, Groq, Whisper, Supabase, WebSockets*

- Built a full-stack AI recruiting platform for live candidate screening using **Groq Whisper** and **Llama 3.3 70B**; engineered a real-time **WebSocket PCM audio pipeline** achieving **sub-2s** end-to-end voice turnaround across concurrent sessions.
- Designed a multi-table **Supabase schema with RLS**, a 5-dimension AI scoring engine, and recruiter dashboard with live session tracking, eliminating **100% of manual screening scheduling** for beta users.

LedgerLake: Streaming Payments Lakehouse [Github] *Databricks, PySpark, Delta Lake, Spark Structured Streaming*

- Built a **Bronze/Silver/Gold Delta Lake pipeline** processing **10M+ payment events** with Spark Structured Streaming and automated data quality gates for production-style ingestion and transformation.
- Implemented **quarantine workflows, fraud analytics, and governed datasets** using Delta Live Tables and Unity Catalog, enabling reliable downstream analytics across streaming payment data.

Entitle: Federal Benefits Navigator [Github] *React, FastAPI, Python, Gemma 4, Ollama*

- Built a **fully offline Python agent** implementing closed-loop eligibility automation (intake → analyze → rank → generate), delivering ranked eligibility checklists across **50 states** with **zero external API calls**.
- Reduced local inference latency by **54% (13 min to under 6 min)** on Apple Silicon by pre-computing FPL thresholds, disabling Gemma 4 hidden thinking tokens, and batching multi-program eligibility into a single model pass.